Integration and Processing of Problem-Oriented Knowledge Based on Evolutionary Procedures

Victoria Bova, Dmitry Zaporozhets and Vladimir Kureichik

Abstract Nowadays integration and processing of domain-specific knowledge is one of the most important tasks of providing access to heterogeneous information from different subject areas and support for interoperability of intelligent information systems aimed at the sharing of data and knowledge on structural and semantic level. The article discusses a modified approach to the problems of integration and knowledge processing, involves a comparison in their automatic ontology using semantic component metrics. The authors propose an evolutionary approach to the problem of integrating multiple ontologies for interoperability and representation of information and knowledge in intelligent information systems. The objectives of integration and processing of knowledge belong to the class of NP-hard optimization problems and can be implemented using genetic algorithms search for optimal solutions. The authors propose a genetic algorithm that is based on the use of analogues to the evolutionary processes of reproduction, crossover, mutation and natural selection. The researchers have conducted a series of experiments to analyze the developed approach. The findings confirmed the theoretical significance and the prospect of this approach, as well as possible to establish the optimal parameters of the algorithm.

Keywords Intelligent information systems · Integration of data and problem-oriented knowledge · Ontology · Semantic model · Genetic algorithms · Genetic operators

V. Bova · D. Zaporozhets · V. Kureichik (✉)
Southern Federal University, Rostov-on-Don, Russia
e-mail: vkur@sfedu.ru

V. Bova
e-mail: vvbova@yandex.ru

D. Zaporozhets
e-mail: elpilasgsm@gmail.com

A. Abraham et al. (eds.), *Proceedings of the First International Scientific Conference "Intelligent Information Technologies for Industry" (IITI'16)*, Advances in Intelligent Systems and Computing 450, DOI 10.1007/978-3-319-33609-1_21

1 Introduction

The development of new approaches and methods of performance, integration and treatment of problem-oriented knowledge is the main direction of development of modern intelligent information systems (IIS) [1–3]. The structure of the IMS includes diverse knowledge base with their own local information models with different standards of description, as well data and knowledge presentation. When you merge them into a global model, it generates many conflicts: the use of different terminology with referring to similar concepts in the domain of IMS; heterogeneity at the level of model specifications and conceptual semantics; identification and conversion of non-uniform data structures and knowledge [4–6].

All this makes the problem of integration in problems of processing domain-specific knowledge rather complex and multi-level. To resolve this problem we should take into account both the structural and syntactical differences in data models and knowledge that generate a schematic heterogeneity, and semantic properties of data objects to ensure semantic interoperability of data and resolution of semantic conflicts. For this reason, the integration of ontologies for the establishment of subsequent interaction information IIS models is an important task.

The paper proposes an evolutionary approach to the problem of integrating multiple ontologies for compatibility and representation data and knowledge in IIS. Such an approach would identify the priority projects semantic data and knowledge to represent them in the model of integration as well as to eliminate duplication and contradictions of entities and relationships at the level of the domain and data objects from the area of integration.

The objectives of integration and knowledge processing belongs to the class of NP-hard optimization problems and can be implemented with genetic algorithms for finding the optimal solutions.

2 Integration Problems of Knowledge in IIS Ontologies

The integration problem of data and knowledge is characterized by a wide variety of productions tasks, approaches and methods used to solve them [1, 3–5]. In general, the problem of integration is the logical association of data belonging to different sources, which provides a unified view of these data and operation.

Paying attention to the heterogeneity of the data, we should clarify this concept. It is not heterogeneous in terms of physical storage (i.e., regardless of their location and method of storage), but in terms of a model of representation—ontological specifications.

The heterogeneity of ontological specifications appears on the level of conceptual modeling and semantics. Differences that create heterogeneity at the level of factor models [1, 5–8]: in the syntax of the language defining the ontological model; in the expressive capacity models; the semantics of the primitives used in the models. Heterogeneity creates differences on the semantic level [5]: in the names of concepts

and relationships; in approaches to the definition of concepts; in the partition on the domain concepts; in covering the subject area; in the points of view on the subject area. Accordingly, we have a problem of coordination for ontological specifications.

Today there are three main components of the problem of integration of data and knowledge: development of integrated circuits, providing a unified view of the data of different sources based on a unified ontological model; development of mappings between ontological models; development of methods of manipulation, the essence of which is disclosed below.

To solve the problem of semantic heterogeneity of data and knowledge, as well an access to heterogeneous information from different subject areas we propose a modified approach to the construction of the resulting ontologies for multiple source-level matching of concepts, relationships, and attributes. Integration model of heterogeneous IIS data and knowledge reduces to the construction of maps and establish relationships in a unified ontological model based on ontology matching multiple levels of IP model and conceptual semantics. The complexity of the model is selected based on the need for its expressive possibilities and a given set of semantic dependencies.

3 Modified Approach to the Integration Problem

One of the objectives of overcoming the lexical and semantic heterogeneity of IIS data and knowledge is to develop new approaches for sharing and support of ontologies developed independently of each other [2, 6]. We propose the following modified approach for solving the integration problem of IIS data and knowledge, as well providing an access to heterogeneous information from various subject areas. Especially suitable to the construction of the resulting ontology to multiple source is to provide consistency in the structural and semantic levels. The proposed approach to building a model of integration of multiple ontologies does not require replacement of individual single ontology, which is the result of combining. The integration process is understood as a process of establishing a non-uniform mapping of ontologies level of compliance with the possibility to extend the set of operations (methods of manipulation) over them in a semantically significant level. Such an approach would identify the priority projects semantic data and knowledge to represent them in the model of integration as well as to eliminate duplication and contradictions of entities and relationships at the level of the domain and data objects from the area of integration.

We need to solve the following problems for the integration of ontologies for matching application contexts [5–8]:

- binding specifications of information sources with the specifications of the domain to reflect its implicit semantics;
- reduction of different ontologies formalisms to one for comparison of ideas about the subject area;

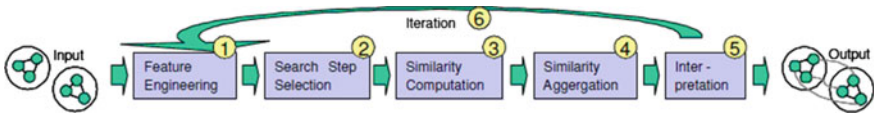

Fig. 1 Process of ontology mapping

- mapping of ontological contexts in one formalism to align them;
- semantic linking of object diagrams elements of information sources and the task based on ontological concepts connection.

The authors propose an algorithm to solve this problem. It includes six operations running consecutively to display the ontology (Fig. 1).

1. Feature Engineering—is relay function of ontologies (ontology analysis of the elements), i.e., converting one format of presentation of the initial ontology, usually result in the format RDF (S), as it is considered standard when working with ontologies.
2. Selection of Text Search Steps is selection of the next step to find a candidate. Choosing an expert semantic proximity search algorithm and semantic distance between pairs of concepts, depending on the goal.
3. Similarity Computation is calculation of similarity, determination of similarity among pairs of ontologies' concepts. It is calculated in step of comparison among ontologies.
4. Similarity Aggregation is aggregation of similarities, i.e., association of entities into one total value confirmation for mapping of connections. Among the pairs of matched entities, we choose one that has a measure of semantic similarity is the highest. Similarity limit is chosen heuristically.
5. Interpretation is formation of the mapping between ontology elements based on similarities, in other words, it is comparison of the names of concepts, assigning the selected class name synthesized from the other two.
6. Iteration is repeating of several steps of the algorithm consists of several stages and stops when it cannot find the new maps.

The proximity and coherence are the main criteria of mapping algorithm of ontologies integral elements.

Evaluation model of semantic proximity for ontologies elements was presented in [6].

4 Evolutionary Approach to the Problem of Knowledge Integration

Hybrid measure takes into account the differences between the compared objects on various grounds and is determined by the dominant value (a measure of Euclid), which allows you to increase the weight of the measures that are important and

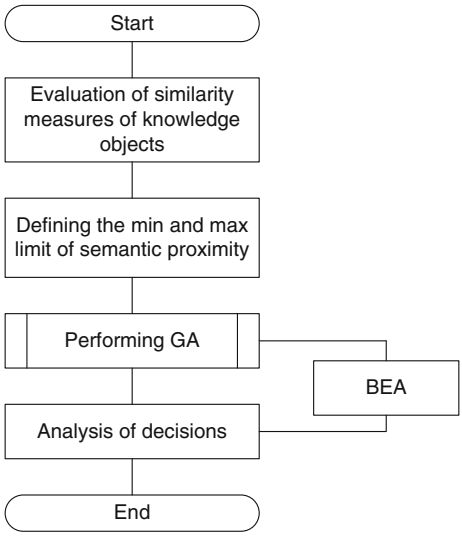

Fig. 2 Generalized architecture of genetic search

virtually ignore measures with small values [9–12]. In this regard, the most promising approach to determine a quantitative measure of proximity is the automatic determination of the weighting coefficients using GA. Generalized structure of the genetic search of weighting coefficients shown in Fig. 2.

The process of genetic research is a sequence of transformations of a finite set of alternatives to another using the mechanisms and principles of genetics and evolution of wildlife [13, 14].

The weighting factors *t*, *r*, *a* allow you to adjust the process of calculating of the semantic proximity of the two concepts. According to the formula

$$\mathrm{C}(\mathrm{k}_i, \mathrm{k}_j) = (\mathrm{t} \cdot \mathrm{C}^{T^{\mathrm{ax}}}(\mathrm{k}_i, \mathrm{k}_j) + \mathrm{r} \cdot \mathrm{C}^{R^{\mathrm{el}}}(\mathrm{k}_i, \mathrm{k}_j) + \mathrm{a} \cdot \mathrm{C}^{A^{\mathrm{tr}}}(\mathrm{k}_i, \mathrm{k}_j)) \quad (1)$$

the evaluating task of concepts semantic proximity of ontology has some limitations: $minf_{t,r,a}(\bar{x}), \bar{x} = (t, r, a) \in F \subseteq S;\ t, r, a \in [0; 1];\ t + r + a = 1$ where $\bar{x}$ is an admissible solution, F is the range of permissible values, and S is the search area.

The objective function (OF) is based on the search for Euclidean distance and has the form:

$$f_{t,r,a} = \sum_{k_i \in ONT, k_j \in ONT'} \left(t * C^{Tax}(k_i, k_j) + r * C^{Rel}(k_i, k_j) + a * C^{Atr}(k_i, k_j) - 1\right)^2 \quad (2)$$

Thus, the search for the weighting factors is to perform four steps.

Evaluation of the similarity of data and knowledge integral ontologies produced in the first stage. The method of calculating the semantic proximity of concepts allows quantifying the similarity between the concepts [6, 14]. We define a values measure limit of proximity to rank the elements of the result set (Fig. 3).

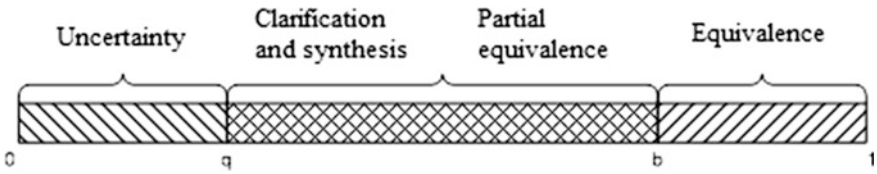

Fig. 3 Levels of threshold values for semantic proximity measures

The parameter b is the minimum limit, which determines the value where it is possible to map the full ontology. The limit where the concepts are accepted in part equivalent is calculated as follows:

$$q = \max(\mathrm{C}\ (\mathrm{k}_i, \mathrm{k}_j)\ |\forall \mathrm{k}_i \in \mathrm{ONT}, \forall \mathrm{k}_j \in \mathrm{ONT}^{'}) \times p_2/100, \tag{3}$$

where p_2 is the percentage where q is taken as similarity limit for establishing a partial equivalence concepts.

The concepts are different in the case when the measure value of semantic proximity is less than the threshold q.

Further GA runs to find the weighting limit and compare with limit values. Block of evolutionary adaptation (BEA) has been introduced to implement feedback search architecture, which is based on interaction with the external environment (the decision-maker) manages the process of finding and configuring the GA [13]. It affects the rearrangement of the current population of alternative solutions and the creation of a new population.

5 GA of Weighting Coefficients Search

The block diagram of GA is shown in Fig. 4. The first stage is the input weighting factors $w_i\{t_i, r_i, a_i\}$, which determine the importance of proximity measures $C^{T^{ax}}(k_i, k_j)$, $C^{R^{el}}(k_i, k_j)$, $C^{A^{tr}}(k_i, k_j)$.

Further, the initial population of alternative solutions P is generated with $t, r, a \in [0; 1]$, $t + r + a = 1$. Each alternative corresponds to the chromosome, which is a $w_i = \langle t_i, r_i, a_i \rangle$, n is the number of chromosomes in the population.

After the step of generation we calculate OF's value for each chromosome according to the formula (2). Selecting pairs of chromosomes for subsequent crossing is made based on the data obtained. Such a choice is made according to the principle of natural selection, where the greatest chance to participate in the creation of new species belongs to chromosomes with the highest values of OF.

The method of the roulette wheel has been chosen for the implementation of operator selection [10]. In spite of the random nature of this procedure, the parent individuals are selected in proportion to the values of OF: each chromosome is

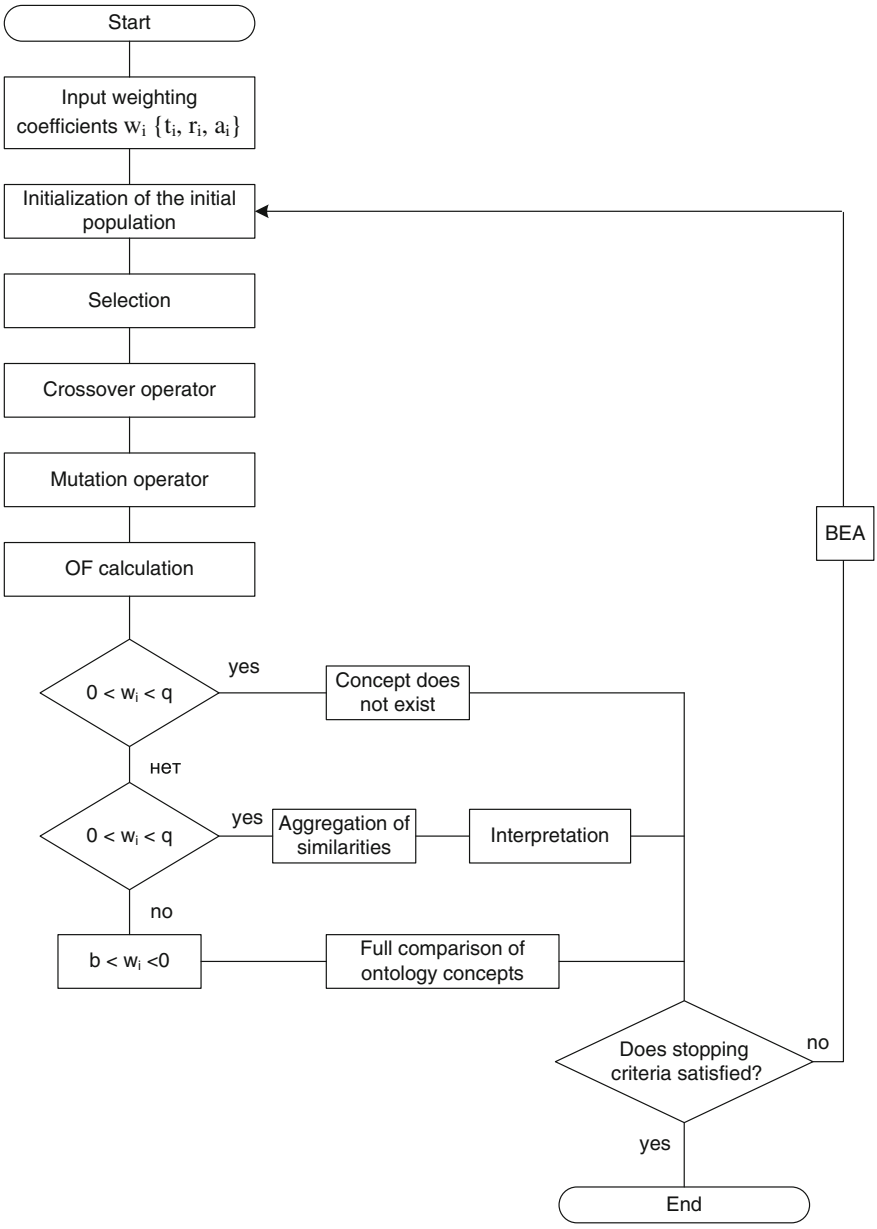

Fig. 4 GA block diagram

associated sector of the roulette wheel, and its value shall be proportional to the value of OF given chromosome, therefore, the greater the value OF, the more sectors on the roulette wheel.

The next step of GA is to perform crossover operator, where its main task is to provide ultimately the most functional features that were present in the set of source solutions.

Ordered crossover operator implemented as a part of the solved problem. Ordered crossover is implemented in transformation stages of genetic material and enables the only real solution.

Break point is selected randomly. Next we make up the left segment of the parent chromosomes P1 in the left segment of the chromosome, the descendant of P1′. The remaining genes P1′ is taken from the second parent chromosome P2 left in an orderly manner. The second descendant P2′ formed in a similar manner.

Break point is selected randomly. Next we make up the left segment of the parent chromosomes P1 in the left segment of the chromosome, the descendant of P1′. The remaining genes P1′ is taken from the second parent chromosome P2 left in an orderly manner. The second descendant P2′ formed in a similar manner.

Gene mutations have the greatest importance in the solutions of the current task. Mutations lead to the appearance of qualitatively new properties of the genetic material. The essence of the developed mutation operator consists of the following. We randomly select a random number in the studied chromosome. Factor mutation determines the intensity of mutations. It determines the fraction of genes undergoing mutations at the current iteration, based on their total amount. If the mutation rate is too small, we get the situation in which a plurality of useful genes simply will not exist in the population. The magnitude selected according to the value after the change of *i*-gene it was in the interval [0; 1].

Using this strategy, the implementation of the operator increases the search space, which is a prerequisite for finding the optimal solution [14].

OF calculation results leads to different operations with ontology concepts presented in the algorithm (Fig. 4).

In this algorithm as a stopping criterion is proposed to use a certain number of iterations. As long as the stopping criterion has not been reached, evolutionary adaptation is performed in the transition to the next iteration.

6 Experimental Results

The problem of integration of ontologies as well as mapping of the ontological concepts involves finding commonalities and differences in the specifications of multiple ontologies for future interoperability and representation of data and problem-oriented knowledge in IIS [5–8, 12].

Ontologies of integral IIS initially have nothing to do; therefore, we need to find semantically similar elements to confirm the correctness of ontologies and semantic relationships established between ontological concepts. Therefore, the purpose of the analysis of algorithms assess the similarity measures we have considered projects that implement the methods of calculating of the semantic proximity of concepts integral ontologies. To evaluate the developed algorithm authors have

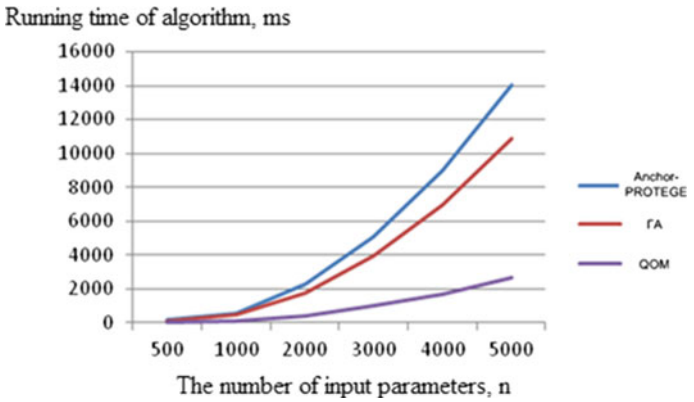

Fig. 5 Graph of dependence of time decisions from the number of input parameters

performed a comparative analysis of such systems using Anchor-PROMPT [5] and QOM [9]. The software supports process of unification and ontology mapping.

The software environment for finding weighting factors has been developed for experimental research work of the proposed GA t, r, a, that allow you to adjust the process of calculating of the semantic proximity of the two concepts. This problem is reduced by solving a system of linear algebraic equations.

The experimental results allowed determining the dependence of the algorithm on the input parameters (weight coefficients): n js the number of chromosomes in the population; chromosome is a tuple $w_i = \langle t_i, r_i, a_i \rangle$.

The graphs of working time of the algorithm, as well as the Anchor-PROMPT and QOM, the number of input data shown in Fig. 5.

The proposed approach presents the original mechanism for mapping and integration of ontologies using a genetic algorithm to determine the proximity of ontology elements according to (weights) copies of concepts. Ontologies are defined as taxonomic concepts with attributes. During the search process, the rules GA matching items ontologies. Precision performance depends on the quality obtained at each iteration GA-effective solutions, weighing the results of determination of similarity of concepts of integral ontologies.

The time complexity of the algorithm is approximately $O(n^2)$.

7 Conclusion

The main advantages of an evolutionary approach to solving the problems of integration and processing of domain-specific knowledge are to identify the key concepts for the construction of the resulting ontology, as well as elimination of subjective descriptions of concepts ontology and depending on perspective development of ontologies. To solve the problem of semantic conflicts we have proposed

the assessment model of semantic proximity based on harmonization of the attribute, the taxonomic and relational similarity measure. Authors have developed GA of determining similarity criterion for the classification of concepts maps in the following groups: equivalence of partial equivalence, summarizing, clarification, uncertainty. The proposed approach allows finding an effective solution to the problem of data integration and knowledge using modified genetic operators and the process of evolutionary adaptation. A distinctive feature of the approach is the automatic calculation of weight coefficients using GA.

Acknowledgment The study was performed by the grant from the Russian Science Foundation (project # 14-11-00242) in the Southern Federal University.

References

1. Bova, V.V., Kureichik, V.V., Legebokov, A.A.: The integrated model of representation model of representation oriented knowledge in information systems In: 8th IEEE International Conference In: Application of Information and Communication Technologies—AICT 2014, pp. 111–115. IEEE Press, Astana, Kazakhstan (2014)
2. Kravchenko, Y.A., Kureichik, V.V.: Knowledge management based on multi-agent simulation in informational systems. In: 8th IEEE International Conference "Application of Information and Communication Technologies—AICT 2014", pp. 264–267. IEEE Press, Astana, Kazakhstan (2014)
3. Bova, V.V., Kravchenko, Y.A., Kureichik, V.V.: Decision support systems for knowledge management. In: Advances in Intelligent Systems and Computing, pp. 123–130. Czech Republic (2015)
4. Rodzin, S.I., Podzina, L.S.: Mobile learning systems and ontology. In: Advances in Intelligent Systems and Computing, pp. 45–54. Czech Republic (2015)
5. Skvortsov, N.A.: Questions harmonize heterogeneous ontological models and ontological context. Ontol. Model. J. 149–166 (2008)
6. Bova, V., Kureichik, V., Zaruba, D.: Data and knowledge classification in intelligence informational systems by the evolutionary method. In: Proceedings of the 6th International Conference Cloud System and Big Data Engineering (Confluence'2016), pp. 6–11. India (2015)
7. Bova, V.V., Kravchenko, Y.A., Kureichik, V.V.: Development of distributed information systems: ontological approach. In: Advances in Intelligent Systems and Computing, pp. 113–122. Czech Republic (2015)
8. Gavrilova, T.A.: The ontological approach to knowledge management in the development of corporate information systems. News Artif. Intell. J. **1**, 24–30 (2003)
9. Tuzovskiy, A.F.: Working with ontologies in the knowledge management system the organization. In: Abstracts of the Second International Conference on Cognitive Science CogSci-2006, pp. 581–583. SPb: SPbGU (2006)
10. Zaporozhets, D.Y., Zaruba, D.V., Kureichik, V.V.: Hybrid bionic algorithms for solving problems of parametric optimization. J. World Appl. Sci. J. **23**, 1032–1036 (2013)
11. Rodzin, S., Rodzina, L.: Theory of bioinspired search for optimal solutions and its application for the processing of problem-oriented knowledge. In: 8th IEEE International Conference "Application of Information and Communication Technologies—AICT 2014", pp. 142–147. IEEE Press, Astana, Kazakhstan (2014)
12. Bova, V.V., Kureichik, V.V., Legebokov, A.A.: Integration of ontologies in scope of model and conceptual semantics: modified approach. In: 9th IEEE International Conference

"Application of Information and Communication Technologies—AICT 2015", pp. 111–115. IEEE Press, Rostov-on-Don, Russia (2015)
13. Gladkov, L.A., Gladkova, N.V., Legebokov, A.A.: Organization of knowledge management based on hybrid intelligent methods. In: Advances in Intelligent Systems and Computing, pp. 107–112. Czech Republic (2015)
14. Bova, V.V., Legebokov, A.A., Gladkov, L.A.: Problem-oriented algorithms of solutions search based on the methods of swarm intelligence. J. World Appl. Sci. J. **27**, 1201–1205 (2013)